HIV Sequence Compendium 2002

Editors

Carla Kuiken
Los Alamos National Laboratory

Brian Foley
Los Alamos National Laboratory

Eric Freed
National Institute of Allergy and Infectious Diseases

Beatrice Hahn
University of Alabama

Preston Marx
ADARC

Francine McCutchan
Henry M. Jackson Foundation

John W. Mellors
University of Pittsburgh

Steven Wolinsky
Northwestern University

Bette Korber
Los Alamos National Laboratory

Project Officer
James Bradac
Division of AIDS
National Institute of Allergy and Infectious Diseases

Los Alamos Database and Analysis Staff
Werner Abfalterer, Charles Calef, Brian Gaschen,
Kristina Kommander, Dorothy Lang, Catherine Miller, Ming Zhang

This publication is being funded by the Division of AIDS, National Institute of Allergy and Infectious Diseases, through an interagency agreement with the U.S. Department of Energy.

Published by Theoretical Biology and Biophysics
Group T-10, Mail Stop K710
Los Alamos National Laboratory, Los Alamos, New Mexico 87545 U.S.A.

LA-UR 03-3564

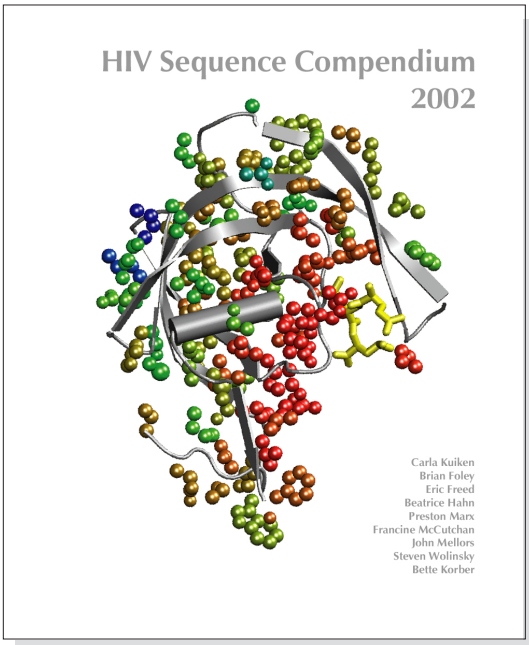
http://hiv-web.lanl.gov

CONTENTS

ACKNOWLEDGMENTS

The HIV Sequence Database and Analysis Project is funded by the Vaccine and Prevention Research Program of the AIDS Division of the National Institute of Allergy and Infectious Diseases (Dr. James Bradac, Project Officer) through an interagency agreement with the U.S. Department of Energy.

The Cover



The cover illustration of this year's HIV Sequence Compendium depicts the HIV-1 protease enzyme structure of the 1A30 PDB Database entry rendered with Visual Molecular Dynamics software: http://www.ks.uiuc.edu/Research/vmd. This is one of the tools discussed in the review article "Visualizing Lentiviral Protein 3D Structures" on page 89 of this volume. Only chain A, one of the homodimer protease molecules, and chain C, a tripeptide inhibitor, are shown. Chain A amino acid side chains are colored by amino acid conservation score (red most conserved through green to blue most variable) calculated from a multiple sequence alignment of lentiviral protease protein sequences using the ConSurf server: http://bioinfo.tau.ac.il/~consurf/results/1054247768/index.htm. It can be seen that amino acids surrounding the enzyme's active site are highly conserved. The chain A backbone is rendered in cartoon mode in gray, and the chain C peptide inhibitor is shown in bond mode and colored yellow. The rendering was created by Brian Foley with assistance from Fabian Glaser at the ConSurf server and Ben McMahon at Los Alamos.

Citing this publication

This publication should be cited as *HIV Sequence Compendium 2002*, Kuiken C, Foley B, Freed E, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S, and Korber B, editors. Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, LA-UR number 03-3564.

Introduction

This compendium is an annual printed summary of the data contained in the HIV sequence database. In these compendia we try to present a judicious selection of the data in such a way that it is of maximum utility to HIV researchers. Traditionally, we present the sequence data themselves in the form of alignments:

Section II, an alignment of a selection of HIV-1/SIVcpz full-length genomes (a lot of LAI-like sequences, for example, have been omitted because they are so similar that they bias the alignment);

Section III, a combined HIV-1/HIV-2/SIV whole genome alignment;

Sections IV–VI, amino acid alignments for HIV-1/SIV-cpz, HIV-2/SIV, and SIVagm.

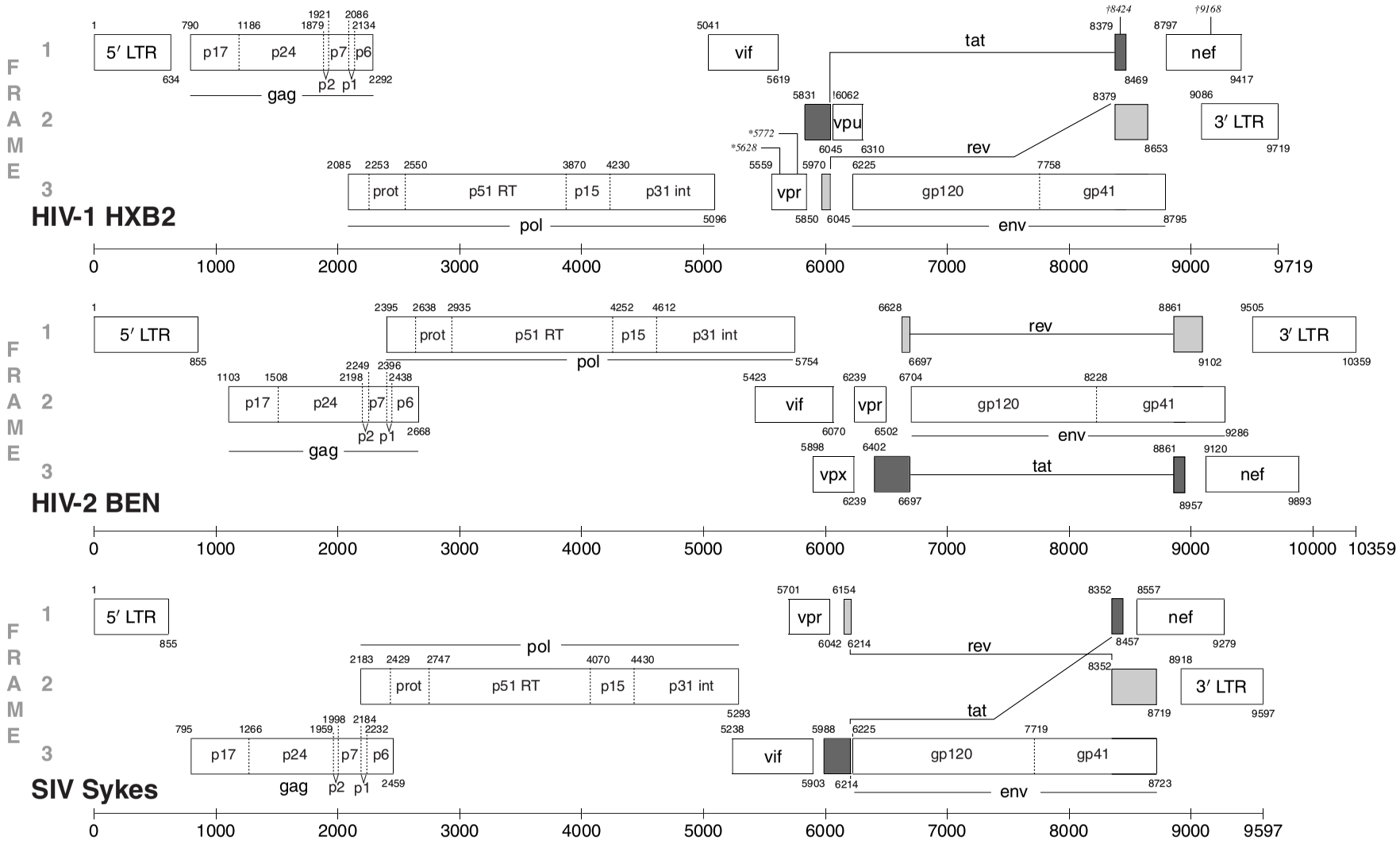
The HIV-2/SIV and SIVagm amino acid alignments are separate because the genetic distances between these groups are so great that presenting them in one alignment would make it very elongated because of the large number of gaps that have to be inserted. As always, tables with extensive background information gathered from the literature accompany the whole genome alignments.

The collection of whole-gene sequences in the database is now large enough that we have abundant representation of most subtypes. For many subtypes, and especially for subtype B, a large number of sequences that span entire genes were not included in the printed alignments to conserve space. A more complete version of all alignments is available on our website, http://hiv-web.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html. Importantly, all these alignments have been edited to include only one sequence per person, based on phylogenetic trees that were created for all of them, as well as on the literature. Because of the number of sequences available, we have decided to use a different selection principle this year, based on the epidemiological importance of the subtypes. Subtypes A–D and CRFs 01 and 02 are by far the most widespread variants, and for these (when available) we have included 8–10 representatives in the alignments. The other subtypes and CRFs are of lesser importance, and of these 4–5 each, or as many as are available, were included.

In the alignments we have also included the 'Circulating Recombinant Forms', mosaic genomes that have epidemiological significance. See the 1999 review of nomenclature (http://hiv-web.lanl.gov/content/hiv-db/REVIEWS/nomenclature/Nomen.html) for more on CRFs, and see for an overview of the patterns of known CRFs. Amino acid alignment chapters begin with an annotation table that includes sequence names, accession numbers, genomic region represented, author, and references. We have made an effort to bring the HIV-2/SIV and SIVagm alignments up-to-date as well.

Reprints of all reviews are available from our website in the form of both HTML and PDF files. As always, we are open to complaints and suggestions for improvement. With the effort that goes into producing these volumes, we sincerely hope they will be widely used by the research community. Inquiries and comments regarding the Compendium should be addressed to:

Dr. Carla Kuiken
Theoretical Division, T-10, MS K710, LANL, Los Alamos, NM 87545
Ph: (505)-665-6463; fax: (505)-665-3493; e-mail: kuiken@t10.lanl.gov



Landmarks of the HIV-1, HIV-2, and SIV genomes. The gene start, indicated by the small number in the upper left corner of each rectangle normally records the position of the `a` in the `atg` start codon for that gene while the number in the lower right records the last position of the stop codon. For *pol*, the start is taken to be the first `t` in the sequence `ttttttag` which forms part of the stem loop that potentiates ribosomal slippage on the RNA and a resulting –1 frameshift and the translation of the gag-pol polyprotein. The *tat* and *rev* spliced exons are shown as shaded rectangles. In HXB2, **5628* and **5772* mark positions of frameshifts in the `vpr` gene; !6062 indicates a defective `acg` start codon in *vpu*; *†8424*, and *†9168* mark premature stop codons in *tat* and *nef*. See Korber et al., Numbering Positions in HIV Relative to HXB2CG, in *Human Retroviruses and AIDS, 1998* p. 102. Available from `http://hiv-web.lanl.gov/HTML/reviews/HXB2.html`.

HIV/SIV PROTEINS

NAME	SIZE	FUNCTION	LOCALIZATION
Gag MA	p17	membrane anchoring; env interaction; nuclear transport of viral core. (myristylated protein)	virion
CA	p24	core capsid	virion
NC	p7	nucleocapsid, binds RNA	virion
	p6	binds Vpr	virion
Protease (PR)	p15	gag/pol cleavage and maturation	virion
Reverse transcriptase (RT), RNase H	p66 p51	reverse transcription, RNase H activity	virion
Integrase (IN)		DNA provirus integration	virion
Env	gp120 gp41	external viral glycoproteins bind to CD4 and secondary receptors	plasma membrane, virion envelope
Tat	p16/p14	viral trancriptional transactivator	primarily in nucleolus/nucleus
Rev	p19	RNA transport, stability and utilization factor (phosphoprotein)	primarily in nucleolus/nucleus shuttling between nucleolus and cytoplasm
Vif	p23	promotes virion maturation and infectivity	cytoplasm (cytosol, membranes) virion
Vpr	p10-15	promotes nuclear localization of preintegration complex, inhibits cell division, arrests infected cells at G2/M	virion, nucleus (nuclear membrane?)
Vpu	p16	promotes extracellular release of viral particles; degrades CD4 in the ER; (phosphoprotein only in HIV-1 and SIVcpz)	integral membrane protein
Nef	p27-p25	CD4 and class I downregulation (myristylated protein)	plasma membrane, cytoplasm (virion?)
Vpx	p12-16	vpr homolog (not in HIV-1, only in HIV-2 and SIV)	virion (nucleus?)

LANDMARKS:

HIV GENOMIC STRUCTURAL ELEMENTS

LTR Long terminal repeat, the DNA sequence flanking the genome of integrated proviruses. It contains important regulatory regions, especially those for transcription initiation and polyadenylation.

TAR Target sequence for viral transactivation, the binding site for Tat protein and for cellular proteins; consists of approximately the first 45 nucleotides of the viral mRNAs in HIV-1 (or the first 100 nucleotides in HIV-2 and SIV.) TAR RNA forms a hairpin stem-loop structure with a side bulge; the bulge is necessary for Tat binding and function.

RRE Rev responsive element, an RNA element encoded within the env region of HIV-1. It consists of approximately 200 nucleotides (positions 7327 to 7530 from the start of transcription in HIV-1, spanning the border of gp120 and gp41). The RRE is necessary for Rev function; it contains a high affinity site for Rev; in all, approximately seven binding sites for Rev exist within the RRE RNA. Other lentiviruses (HIV-2, SIV, visna, CAEV) have similar RRE elements in similar locations within env, while HTLVs have an analogous RNA element (RXRE) serving the same purpose within their LTR; RRE is the binding site for Rev protein, while RXRE is the binding site for Rex protein. RRE (and RXRE) form complex secondary structures, necessary for specific protein binding.

CRS Cis-acting repressive sequences postulated to inhibit structural protein expression in the absence of Rev. One such site was mapped within the pol region of HIV-1. The exact function has not been defined; splice sites have been postulated to act as CRS sequences.

INS Inhibitory/Instability RNA sequences found within the structural genes of HIV-1 and of other complex retroviruses. Multiple INS elements exist within the genome and can act independently; one of the best characterized elements spans nucleotides 414 to 631 in the gag region of HIV-1. The INS elements have been defined by functional assays as elements that inhibit expression posttranscriptionally. Mutation of the RNA elements was shown to lead to INS inactivation and up regulation of gene expression.

GENES AND GENE PRODUCTS

GAG The genomic region encoding the capsid proteins (group specific antigens). The precursor is the p55 myristylated protein, which is processed to p17 (MAtrix), p24 (CApsid), p7 (NucleoCapsid), and p6 proteins, by the viral protease. Gag associates with the plasma membrane where the virus assembly takes place. The 55 kDa Gag precursor is called assemblin to indicate its role in viral assembly.

POL The genomic region encoding the viral enzymes protease, reverse transcriptase and integrase. These enzymes are produced as a Gag-pol precursor polyprotein, which is processed by the viral protease; the Gag-pol precursor is produced by ribosome frameshifting at the C-terminus of gag.

ENV Viral glycoproteins produced as a precursor (gp160) which is processed to give a noncovalent complex of the external glycoprotein gp120 and the transmembrane glycoprotein gp41. The mature gp120-gp41 proteins are bound by non-covalent interactions and are associated as a trimer on the cell surface. A substantial amount of gp120 can be found released in the medium. gp120 contains the binding site for the CD4 receptor, and the seven transmembrane domain chemokine receptors that serve as co-receptors for HIV-1.

TAT Transactivator of HIV gene expression. One of two essential viral regulatory factors (Tat and Rev) for HIV gene expression. Two forms are known, Tat-1 exon (minor form) of 72 amino acids and Tat-2exon (major form) of 86 amino acids. Low levels of both proteins are found in persistently infected cells. Tat has been localized primarily in the nucleolus/nucleus by immunofluorescence. It acts by binding to the TAR RNA element and activating transcription

initiation and/or elongation from the LTR promoter. It is the first eukaryotic transcription factor known to interact with RNA rather than DNA and may have similarities with prokaryotic anti-termination factors. Extracellular Tat can be found and can be taken up by cells in culture.

REV The second necessary regulatory factor for HIV expression. A 19 kD phosphoprotein, localized primarily in the nucleolus/nucleus, Rev acts by binding to RRE and promoting the nuclear export, stabilization and utilization of the viral mRNAs containing RRE. Rev is considered the most functionally conserved regulatory protein of lentiviruses. Rev cycles rapidly between the nucleus and the cytoplasm.

VIF Viral infectivity factor, a basic protein of typically 23 kD. Promotes the infectivity but not the production of viral particles. In the absence of Vif the produced viral particles are defective, while the cell-to-cell transmission of virus is not affected significantly. Found in almost all lentiviruses, Vif is a cytoplasmic protein, existing in both a soluble cytosolic form and a membrane-associated form. The latter form of Vif is a peripheral membrane protein that is tightly associated with the cytoplasmic side of cellular membranes. Some recent observations suggest that Vif functions late in replication to modulate assembly, budding, and/or maturation the N-terminal half of Vif (N'-Vif) specifically interacts with viral protease.

VPR Vpr (viral protein R) is a 96-amino acid (14 kd) protein, which is incorporated into the virion. It interacts with the p6 gag part of the Pr55 gag precursor. Vpr detected in the cell is localized to the nucleus. Proposed functions for Vpr include the targeting the nuclear import of preintegration complexes, cell growth arrest, transactivation of cellular genes, and induction of cellular differentiation. It is found in HIV-1, HIV-2, SIVmac and SIVmnd. It is homologous to the vpx protein.

VPU Vpu (viral protein U) is unique to HIV-1 and SIVcpz, a close relative of HIV-1. There is no similar gene in HIV-2 or other SIVs. Vpu is a 16-kd (81-amino acid) type I integral membrane protein with at least two different biological functions: (a) degradation of CD4 in the endoplasmic reticulum, and (b) enhancement of virion release from the plasma membrane of HIV-1-infected cells. Env and Vpu are expressed from a bicistronic mRNA. Vpu probably possesses an N-terminal hydrophobic membrane anchor and a hydrophilic moiety. It is phosphorylated by casein kinase II at positions Ser52 and Ser56. Vpu is involved in env maturation and is not found in the virion. Vpu has been found to increase susceptibility of HIV-1 infected cells to Fas killing.

NEF A multifunctional 27-kd myristylated protein produced by an ORF located at the $3'$ end of the primate lentiviruses. Other forms of Nef are known, including nonmyristylated variants. Nef is predominantly cytoplasmic and associated with the plasma membrane via the myristyl residue linked to the conserved second amino acid (Gly). Nef has also been identified in the nucleus and found associated with the cytoskeleton in some experiments. One of the first HIV proteins to be produced in infected cells, it is the most immunogenic of the accessory proteins. The nef genes of HIV and SIV are dispensable *in vitro*, but are essential for efficient viral spread and disease progression *in vivo*. Nef is necessary for the maintenance of high virus loads and for the development of AIDS in macaques, and viruses with defective Nef have been detected in some HIV-1 infected long term survivors. Nef downregulates CD4, the primary viral receptor, and MHC class I molecules, and these functions map to different parts of the protein. Nef interacts with components of host cell signal transduction and clathrin-dependent protein sorting pathways. It increases viral infectivity. Nef contains PxxP motifs that bind to SH3 domains of a subset of Src kinases and are required for the enhanced growth of HIV but not for the downregulation of CD4.

VPX A virion protein of 12 kD found only in HIV-2/SIVmac/SIVsm and not in HIV-1 or SIVagm. This accessory gene is a homolog of HIV-1 vpr, and HIV-2/SIV carry both vpr and vpx. Vpx function in relation to xpr is not fully elucidated; both are incorporated into virions at levels comparable to gag proteins through interactions with Gag p6. Vpx is necessary for efficient replication of SIV in PBMCs. Progression to AIDS and death in SIV-infected animals can occur in the absence of Vpr or Vpx. Double mutant virus lacking both vpr and vpx was attenuated,

whereas the single mutants were not, suggesting a redundancy in the function of Vpr and Vpx related to virus pathogenicity.

STRUCTURAL PROTEINS/VIRAL ENZYMES The products of gag, pol and env genes, which are essential components of the retroviral particle.

REGULATORY PROTEINS Tat and Rev proteins of HIV/SIV and Tax and Rex proteins of HTLVs. They modulate transcriptional and posttranscriptional steps of virus gene expression and are essential for virus propagation.

ACCESSORY OR AUXILIARY PROTEINS Additional virion and non-virion- associated proteins produced by HIV/SIV retroviruses: Vif, Vpr, Vpu, Vpx, Nef. Although the accessory proteins are in general not necessary for viral propagation in tissue culture, they have been conserved in the different isolates; this conservation and experimental observations suggest that their role *in vivo* is very important. Their functional importance continues to be elucidated.

COMPLEX RETROVIRUSES Retroviruses regulating their expression via viral factors and expressing additional proteins (regulatory and accessory) essential for their life cycle.